(19)日本国特許庁（ＪＰ）　(12) 公開特許公報（Ａ）　(11)特許出願公開番号

特開平6－332778

(43)公開日　平成６年(1994)12月２日

(51)Int.Cl.[5]	識別記号	庁内整理番号	ＦＩ	技術表示箇所
Ｇ０６Ｆ　12/00	５３１　Ｒ	8944－5B		

審査請求　有　　請求項の数６　ＯＬ　（全 17 頁）

(21)出願番号　特願平6－86114

(22)出願日　平成６年(1994)４月25日

(31)優先権主張番号　０６６３６０

(32)優先日　1993年５月21日

(33)優先権主張国　米国（ＵＳ）

(71)出願人　390009531
インターナショナル・ビジネス・マシーンズ・コーポレイション
INTERNATIONAL BUSINESS MASCHINES CORPORATION
アメリカ合衆国10504、ニューヨーク州　アーモンク　（番地なし）

(72)発明者　チャンドラセカラン・モーハン
アメリカ合衆国95120　カリフォルニア州　サンノゼ　ポストウッド・ドライブ727

(74)代理人　弁理士　合田　潔　（外２名）

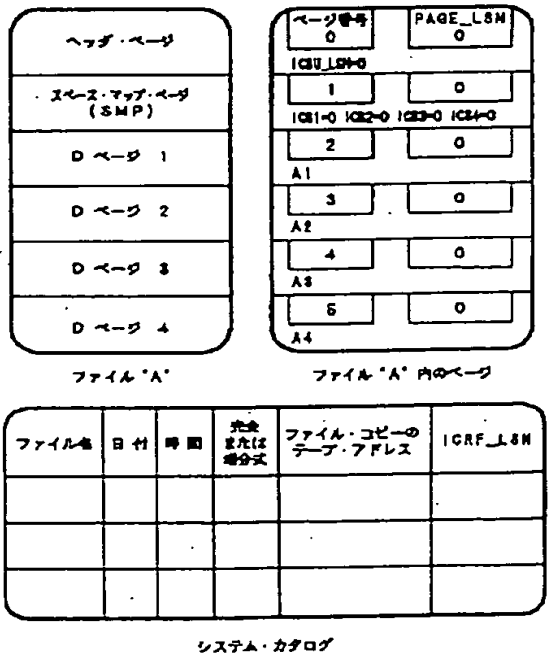
最終頁に続く

(54)【発明の名称】　トランザクション管理方法

(57)【要約】

【目的】　ログ・ベースの段階式コミット・トランザクション管理システム（ＴＭＳ）において、修正可能なページのファイルをアーカイブ保存するための方法を提供すること。

【構成】　これは、１対の大域ログ・シーケンス番号の管理によって達成される。第１の番号（ＩＣＢＵ＿ＬＳＮ）と各データ・ページＬＳＮをページが修正されるときに比較することにより、共通状況ページを、変化した状況を正確に反映するように更新することができる。同じページへのその後の修正には、状況ページの訂正を必要としない。状況ページ標徴は、バックアップ手順の一環として、増分式複写のためにページ・コピー・セットを確認するためにリセットされる。第２の番号（ＩＣＲＦ＿ＬＳＮ）は、ファイルの復元に使用される。この場合、ＩＣＲＦ＿ＬＳＮが、最近のコピーが作成されて以降の、ログ内の再実行のための点を定義する。



論理ファイル，ページおよびシステム・カタログの構成

1

【特許請求の範囲】

【請求項１】ログと、プロセッサと、記憶されたページが前記プロセッサとの間でステージングされる記憶サブシステムとを備えた段階式コミット・トランザクション管理システム（ＴＭＳ）において、ヘッダ・ページと、状況ページ（ＳＭＰ）と、少なくとも１つのデータ・ページとを有する修正可能なページのファイルをアーカイブ保存するための方法であって、

前記プロセッサが各トランザクションに応答して、前記ページのうちの所定のページに対して選択的修正を実行し、ページ状態に対する各修正が前記ログに記録され、昇順ログ：シーケンス番号（ｐａｇｅ＿ＬＳＮ）が割り当てられ、前記ｐａｇｅ＿ＬＳＮが前記ページ上に記録され、（ａ）前記ヘッダ・ページ上の第１の大域的ログ・シーケンス番号（ＩＣＢＵ＿ＬＳＮ）と、それぞれ前記各データ・ページに対応し、前記ページが直前のアーカイブ動作または複写動作以降に修正されたこと示す、前記ＳＭＰ上の複数の状況ビット（ＩＣＢ）とを生成し維持する段階と、（ｂ）前記ＩＣＢＵ＿ＬＳＮを最大値に設定し、第２の大域的ログ・シーケンス番号（ＩＣＲＦ＿ＬＳＮ）を設定し、システム・カタログ内に該ＩＣＲＦ＿ＬＳＮを記録する段階と、（ｃ）前記ＳＭＰ内の前記ＩＣＢを走査することによってアーカイブ保存されるページのセットを確認し、別の形でセットされたサブセットをリセットする段階と、（ｄ）段階（ｃ）で走査されリセットされたサブセットの各ページを順にラッチし、複写し、ラッチ解除することにより、増分式複写のために、完全複写に関するＩＣＢの状態に関係なくすべてのページについて複写動作を実行する段階と、（ｅ）前記ＩＣＢＵ＿ＬＳＮを現end_of_log ＬＳＮの値に設定し、該ＩＣＢＵ＿ＬＳＮを前記ヘッダ・ページに記録する段階と、（ｆ）各ページ修正に応じて、前記ｐａｇｅ＿ＬＳＮが前記ＩＣＢＵ＿ＬＳＮよりも小さい場合に前記ＳＭＰ内に対応ＩＣＢを設定し、該設定動作をＲＥＤＯ＿ｏｎｌｙログ・レコードによって前記ログ内に記録する段階とを含むことを特徴とする方法。

【請求項２】前記ファイル内の各ページに、他の標徴から独立した昇順のページ番号が記録されており、前記システム・カタログが、前記ファイルの各完全コピーまたは増分式コピーに関する情報を記録するエントリを含み、各エントリが、前記記憶サブシステム内の前記ファイルの各アーカイブ・コピーの位置を指すポインタと、請求項１の段階（ｅ）によるend_of_log ＬＳＮにセットされたＩＣＲＦ＿ＬＳＮとを含み、前記方法が、

（ｇ）ログすることなしにデータが前記ファイルにロードされるたびに各ファイルの完全コピーを作成し、該完全コピーが作成されるまでデータ・ページの修正を中断し、前記システム・カタログ内に前記ポインタ位置とコピー・タイプ（完全）とＩＣＲＦ＿ＬＳＮとを記録する段階と、（ｈ）前記システム・カタログへのエントリを

2

含む各ファイルの増分式コピーまたは完全コピーを予定に従ってまたは都合に合わせて作成する段階と、（ｉ）前記記憶サブシステム内の前記ファイルが使用できない状態の場合、前記システム・カタログによって定義されるような完全コピーおよび増分式コピーの内容をページ番号に従ってマージすることによって前記ファイルを回復し、マージされたページ・セットを、前記ＩＣＲＦ＿ＬＳＮよりも大きいまたは等しいＬＳＮを有する前記ログ・レコードに従って修正する段階と、

を含むことを特徴とする、請求項１に記載の方法。

【請求項３】前記ログ内に記録された動作が、ＲＥＤＯ／ＵＮＤＯタイプのものとＲＥＤＯ＿ｏｎｌｙログ・レコードタイプのものとを含み、前記ヘッダ・ページに対するＩＣＢＵ＿ＬＳＮ修正が、ＲＥＤＯ＿ｏｎｌｙログ・レコードを介して前記ログ内に記録されることを特徴とする、請求項１に記載の方法。

【請求項４】前記ＴＭＳが、すでに修正済みのまたは現在修正中のページのリストを保持し、前記方法の諸段階が、（ａ’）ＩＣＢＵ＿ＬＳＮを最大値に設定し、修正されたリスト内に各ページをＳラッチする段階と、（ｂ’）前記ＩＣＢＵ＿ＬＳＮを非最大値に設定し、ＩＣＲＦ＿ＬＳＮの値を選択する段階と、（ｃ’）前記ＳＭＰ内のＩＣＢを走査することによってアーカイブ保存されるページのセットを確認し、別の形でセットされたサブセットをリセットする段階と、（ｄ’）段階（ｃ）で走査されリセットされたサブセットの各ページを順にラッチし、複写し、ラッチ解除することにより、増分式複写のために、完全複写に関するＩＣＢ状態に関係なくすべてのページについて複写動作を実行する段階とを含むように修正されることを特徴とする、請求項１に記載の方法。

【請求項５】ログと、プロセッサと、記憶されたページが前記プロセッサとの間でステージングされる記憶サブシステムとを備えた段階式コミット・トランザクション管理システム（ＴＭＳ）において、修正可能なページのデータベースをアーカイブ保存するための方法であって、

前記プロセッサが各トランザクションに応答して、ページのうちの所定のページに対して選択的修正を実行し、各ページ状態に対する各修正が前記ログに記録され、昇順ログ・シーケンス番号（ｐａｇｅ＿ＬＳＮ）が割り当てられ、前記ページＬＳＮが前記ページ上に記録され、（ａ）（１）ヘッダ・ページと、少なくとも１つのデータ・ページと、前の複写動作の実行以降に前記データ・ページが修正されたかどうかを示す各データ・ページに関する状況ビット（ＩＣＢ）を含む、少なくとも１つのスペース・マップ・ページ（ＳＭＰ）とを備える区画に関する、データベース制御ブロック（ＤＢＣＢ）を呼び出し、（２）前記ＤＢＣＢ内の第１の種類のＬＳＮ（ＩＣＢＵ＿ＬＳＮ）を最大値に原子的に設定し、第２の種

類のＬＳＮ（ＩＣＲＦ＿ＬＳＮ）を現end_of_log ＬＳＮによって設定された値に設定し、システム・カタログ内に前記ＩＣＲＦ＿ＬＳＮを記録し、（３）最終複写動作以降に更新されたことを示すデータ・ページのすべてのＩＣＢをリセットすることによって各ＳＭＰについて複写されるデータ・ページのセットを確認し、前記ログ内にリセットされたすべてのＩＣＢを記録することによって、前記データベースの前記区画に対する複写動作を初期設定する段階と、

（ｂ）（１）各ＳＭＰについて、段階（ａ）（３）で識別された各データ・ページをラッチし、複写し、ラッチ解除し、（２）前記ＤＢＣＢおよび前記ヘッダ・ページ内の前記ＩＣＢＵ＿ＬＳＮを現end_of_log ＬＳＮ値に更新し、（３）動作をコミットすることによって、前記複写動作を実行する段階と、

（ｃ）各ページ修正に応じて、前記ｐａｇｅ＿ＬＳＮが前記ＩＣＢＵ＿ＬＳＮよりも小さい場合に状況ページ内に対応ＩＣＢを設定し、該設定動作を前記ログ内に記録する段階とを含むことを特徴とする方法。

【請求項６】ログと、プロセッサと、記憶されたページが前記プロセッサとの間でステージングされる記憶サブシステムとを備えた段階式コミット・トランザクション管理システム（ＴＭＳ）において、ヘッダ・ページと、状況ページ（ＳＭＰ）と、少なくとも１つのデータ・ページとを有する修正可能なページのファイルをアーカイブ保存するための方法であって、前記ファイル内の各ページに他の標徴から独立して昇順ページ番号が記録されており、

前記プロセッサが各トランザクションに応答して、前記ページのうちの所定のページに対して選択的修正を実行し、ページ状態に対する各修正が前記ログに記録され、昇順ログ・シーケンス番号（ｐａｇｅ＿ＬＳＮ）が割り当てられ、前記ｐａｇｅ＿ＬＳＮが前記ページ上に記録され、（ａ）前記ヘッダ・ページ上の第１の大域的ログ：シーケンス番号（ＩＣＢＵ＿ＬＳＮ）と、それぞれ各データ・ページに対応し、前記ページが直前のアーカイブ動作または複写動作以降に修正されたことを示す、前記ＳＭＰ上の複数の状況ビット（ＩＣＢ）とを生成し維持する段階と、（ｂ）前記ＩＣＢＵ＿ＬＳＮを最大値に設定し、第２の大域的ログ・シーケンス番号（ＩＣＲＦ＿ＬＳＮ）を設定し、前記ファイルの各完全コピーまたは増分式コピーに関する情報を記録するエントリを含み、各エントリが、前記記憶サブシステム内の前記ファイルの各アーカイブ・コピーの位置を指すポインタと、end_of_log ＬＳＮに設定されたＩＣＲＦ＿ＬＳＮとを含む、システム・カタログ内に前記ＩＣＲＦ＿ＬＳＮを記録する段階と、（ｃ）前記ＳＭＰ内の前記ＩＣＢを走査することによってアーカイブ保存されるページのセットを確認し、別の形でセットされたサブセットをリセットする段階と、（ｄ）段階（ｃ）で走査されリセットされたサブセットの各ページを順にラッチし、複写し、ラッチ解除することにより、増分式複写のために、完全複写に関するＩＣＢの状態に関係なくすべてのページについて複写動作を実行する段階と、（ｅ）前記ＩＣＢＵ＿ＬＳＮを現end_of_log ＬＳＮ値に設定し、該ＩＣＢＵ＿ＬＳＮを前記ヘッダ・ページに記録する段階と、

（ｆ）各ページ修正に応じて、前記ｐａｇｅ＿ＬＳＮが前記ＩＣＢＵ＿ＬＳＮよりも小さい場合に前記ＳＭＰ内に対応ＩＣＢを設定し、該設定動作をＲＥＤＯ＿ｏｎｌｙログ・レコードによって前記ログ内に記録する段階と、（ｇ）ログすることなしにデータが前記ファイルにロードされるたびに各ファイルの完全コピーを作成し、該完全コピーが作成されるまでデータ・ページの修正を中断し、前記システム・カタログ内に前記ポインタ位置とコピー・タイプ（完全）とＩＣＲＦ＿ＬＳＮとを記録する段階と、（ｈ）前記システム・カタログへのエントリを含む各ファイルの増分式コピーまたは完全コピーを予定に従ってまたは都合に合わせて作成する段階と、

（ｉ）前記記憶サブシステム内の前記ファイルが使用できない状態の場合、前記システム・カタログによって定義されるような完全コピーおよび増分式コピーの内容をページ番号に従ってマージすることによって前記ファイルを回復し、マージされたページ・セットを、前記ＩＣＲＦ＿ＬＳＮよりも大きいまたは等しいＬＳＮを有する前記ログ・レコードに従って修正する段階と、

を含む方法。

【発明の詳細な説明】

【０００１】

【産業上の利用分野】本発明は、複写によってアプリケーションの実行が妨げられない、完全なまたは増分式のバックアップ複写による、情報操作システムにおけるデータの保存に関する。より詳細には、本発明はログ・ベースの段階式コミット・トランザクション管理システム（ＴＭＳ）において、修正可能なページのファイルをアーカイブ保存するための方法に関する。

【０００２】

【従来の技術】以下の説明は、多段階ログ・ベース・トランザクション管理システムの諸態様、およびそうしたシステムにおいて修正されたデータ・ページをアーカイブ保存することを対象とする。

【０００３】トランザクション管理システムの諸態様

ＴＭＳは一般に、ＣＰＵと、修正可能なページのデータベースを記憶する外部ＤＡＳＤサブシステムと、ページにアクセスしそれを記憶装置とＣＰＵの間で転送するための手段を含む。ＣＰＵで実行されるアプリケーションは、「トランザクション」を通じてページと対話する。「トランザクション」とは、回復可能な資源の一貫状態を、全ての中間点で必ずしも一貫性を保存することなく、別の一貫状態に変形させる一連の動作を含む、論理作業単位と定義される。すなわち、トランザクションと

5

は、データベース内の選択されたページをある状態から別の状態に原子的に変化させる有界の一連の動作である。したがって各トランザクションは、完了するか、または打ち切らなければならず、中間状態は許されない。

【０００４】ログ・ベースのＴＭＳでは、ページに対するすべての変更は、ＲＥＤＯレコードおよびＵＮＤＯレコードの形でログに書き込まれる。物理的ログは、ＤＡＳＤの予約部分やテープ・サブシステムなどでよい。ここで、ＲＥＤＯレコードまたはＵＮＤＯレコード自体が、変更または修正されたページをそのＤＡＳＤ記憶位置に書き戻す前に、ログに書き込まれる。これをライト・アヘッド・ロギングと言う。これらの変更レコードは、完了点まで進行したトランザクションを再作成またはＲＥＤＯするために、あるいは完了点まで進行していなかったトランザクションをロールバックまたはＵＮＤＯするために使用される。

【０００５】ＴＭＳは、各トランザクションが、ＢＥＧＩＮプリミティブと、ＣＯＭＭＩＴプリミティブまたはＡＢＯＲＴプリミティブ、およびＥＮＤプリミティブによって境界を区切られることを特徴とする。すなわち、トランザクションは次の形である。

"BEGIN ops,ops, ..,ops COMMIT/ABORT ..END"

【０００６】マルチタスク処理が可能なシステムなどでは、あるトランザクションは開始されたばかりで、他のものは「飛行中」で、さらに他のものはコミットされまたは打ち切られており、さらに他のものはついさっき終わったばかりというように、複数のトランザクションが完了時間が重なり合う。これは、トランザクションがデータベース内の選択されたページに対して何らかの更新を実行し、トランザクションがその正常終了（ＥＮＤ）または一貫性中間点（ＣＯＭＭＩＴ／ＡＢＯＲＴ）に到達する以前に障害が発生するとき、それらの更新がＵＮＤＯされることを段階式コミット・システムが保証することを意味する。もしそうでなければ、正常終了（ＥＮＤ）または一貫性中間点（ＣＯＭＭＩＴまたはＡＢＯＲＴ）に到達したトランザクションはＲＥＤＯされる（ＣＯＭＭＩＴまたはＡＢＯＲＴを含む期間中でＥＮＤの直前に）。

【０００７】回復動作を遂行するためには、修正されたページと対応ログ・レコードとの間で情報の同時性を提供しなければならない。これは、修正された各ページに固有な昇順のログ・シーケンス番号（ＬＳＮ）を含め、ログに書き込まれるＲＥＤＯ／ＵＮＤＯレコードにページ番号を含めることによって達成される。したがって、修正されたページがログ内のその変更レコード位置をインデックスし、ログ・レコードが修正されたページをインデックスする。

【０００８】これらの態様の詳細な実施例は、米国特許第５０４３８６６号、および第４５０７７５１号明細書に出ている。

6

【０００９】米国特許第５０４３８６６号明細書は、ＴＭＳにおいて順方向または逆方向の回復中に使用しなければならないログの範囲を定義する１対のログ・シーケンス番号が、周期的に決定されチェックポイント情報としてログ内に記憶される場合、回復時に、チェックポイントを検索し、ログの分析パスを回避するために回復アルゴリズム中で用いられるその構成要素間の機能比較を行うことを教示している。

【００１０】米国特許第４５０７７５１号明細書は、同期点（ＢＥＧＩＮ、ＣＯＭＭＩＴ、ＡＢＯＲＴ）の発生時に、バッファ内容を第２のログに書き出すことによって対処される、揮発性メモリまたはバッファ内に記憶されたデータのジャーナル処理（ロギング）を開示している。障害が発生する場合、第２のログのロールバック（ＲＥＤＯ／ＵＮＤＯ）処理によってバッファの前の状態が再確立され、そうでない場合は第１のログが使用される。

【００１１】アーカイブ保存

当技術分野では、アーカイブ保存とは、冗長性と機密保護の現情報に関して不定期にアクセスされるシステムの一部分にページを複写するという、体系的な記憶管理の慣行を指す。アーカイブ保存またはバックアップ・コピーは経時的に完全または増分式に行われ、かつ予定に従ってまたは都合に合わせて行われる。明らかに、増分式複写は完全複写よりも消費資源が少ない。

【００１２】ＩＢＭテクニカル・ディスクロージャ・ブルテン、Ｎｏ．２５、ｐｐ．３７３０～３７３２（１９８２年１２月発行）に所載のクラス(Crus)他の論文"Incremental Database Log Image Copy"は、スペース・マップ・ページ（ＳＭＰ）と名付けられた共通の状況構造を走査し、最後の複写動作以降にページが変更されたことを示す状況ビット（増分式状況変化ビット（ＩＣＢ）と呼ぶ）を有するデータ・ページだけを複写することによって、データベース内で修正されたページを増分式に複写する方法を開示している。より詳細には、複写動作の後でデータ・ページが最初に修正されるとき、データ・ページ中およびスペース・マップ・ページ中でＩＣＢがセットされる。データ・ページが修正されており、そのＩＣＢがすでにセットされている場合は（最後の複写動作以降の修正を示す）、ＳＭＰのＩＣＢを修正するために何も行われない。ＩＣＢをそのページに置く理由は、そうすると、データ・ページの更新ごとにＳＭＰを訪問するというオーバヘッドが減少するからである。複写の目的で、ＳＭＰが走査され、そのＩＣＢがセットされている各ページが複写され、次いでデータ・ページおよびＳＭＰ上の対応ＩＣＢがリセットされる。

【００１３】ＩＢＭ関係データベース・システムＤＢ２では、修正されたページのディスクへの書込みは、「即時コミット」を有効にするため、ＩＣＢビットをリセットした直後に実行される。これは、ＤＡＳＤ書込み動作

のバッチ処理ではなく、各ページごとに個別に実行されるので、かなりのオーバーヘッド・コストになる。

【００１４】

【発明が解決しようとする課題】本発明の目的は、ログ・ベースの段階式コミットＴＭＳのデータベース中の修正されたページの増分式複写に伴うオーバヘッドを減少させるための方法および手段を提供することである。

【００１５】本発明の他の目的は、アプリケーションの実行を妨げないログ・ベースの段階式コミットＴＭＳ内で修正可能なページのデータベースをアーカイブ保存するための方法および手段を提供することである。

【００１６】本発明のさらに他の目的は、修正可能なページのデータベースへのアクセスを共有する複数ログ・ベース段階式コミット・トランザクション管理システムで動作可能な方法および手段を提供することである。

【００１７】

【課題を解決するための手段】本発明は、ログ・ベースの段階式コミットＴＭＳの一部として使用可能な構造を利用する。これは、スペース・マップ・ページ（ＳＭＰ）、ページ更新状況ビット（ＩＣＢ）、ヘッダ・ページ、データベース制御ブロック（ＤＢＣＢ）、ログ、２種類のＬＳＮ（ＩＣＢＵ＿ＬＳＮとＩＣＲＦ＿ＬＳＮ）、およびシステム・カタログを含む。ＳＭＰは各データ・ページに関するＩＣＢを含み、ヘッダ・ページはデータベースに関するＩＣＢＵ＿ＬＳＮを格納し、ＤＢＣＢは仮想記憶域中のＩＣＢＵ＿ＬＳＮを維持し、システム・カタログはデータベースに関するＩＣＲＦ＿ＬＳＮを格納する。

【００１８】各ＳＭＰは、最後の複写動作以降にｉ番目のページが更新されたかどうかを示す状況ビット（ＩＣＢ）を含む、複数のデータ・ページに関する割当て状況を追跡する。ＩＣＢは、複写動作が実行された後で最初にページが更新される時にセットされ、前記動作によってページが複写される時にリセットされる。「ヘッダ・ページ」は、データベースの制御または管理構造である。

【００１９】ＤＢＣＢは、データベースがオープンされる時に仮想記憶域中で生成され維持される。これはデータベースの現状態を反映する主要制御構造であり、本発明の方法における初期設定段階の一部分として含まれる。またこれは、２つのＬＳＮのうちの一方（ＩＣＢＵ＿ＬＳＮ）のためのレポジトリでもある。第１のＬＳＮは、「イメージ複写ビット更新ＬＳＮ」（ＩＣＢＵ＿ＬＳＮ）と呼ばれる。第２のＬＳＮは、「イメージ複写順方向ロールＬＳＮ」（ＩＣＲＦ＿ＬＳＮ）と呼ばれる。

【００２０】ＩＣＢＵ＿ＬＳＮの値により、ＳＭＰ内のどれか所与のデータ・ページに関するＩＣＢをセットする必要があるかどうかが決まる。ＩＣＢＵ＿ＬＳＮの初期値は、データベースがＤＡＳＤ記憶装置に最初にロードされた後の、end_of_logＬＳＮに設定される。

【００２１】ＩＣＲＦ＿ＬＳＮは、最後の完全コピーまたはその後の増分式コピーを再ロードした後でデータベースを回復するためにＲＥＤＯ（ＲＥＤＯ動作）しなければならないログ・レコードを識別するために、ログをそこから走査しなければならないＬＳＮを示す。

【００２２】本発明の方法により、ＩＣＢＵ＿ＬＳＮを２段階で使用することにより、従来の問題点を解決することができる。

【００２３】第１段階では、ＩＣＢＵ＿ＬＳＮが最大値にセットされ、ＳＭＰ中のＩＣＢを調べることによって複写順方向ロールＬＳＮ（ＩＣＲＦ＿ＬＳＮ）とページのコピー・セットの両方が確認される。増分式複写では、コピー・セットは、そのＩＣＢが最後の複写動作以降に修正されたことを示すように変更された、ページのセットである。

【００２４】第２段階では、コピー・セット中のページを複写した後で、ＩＣＢＵ＿ＬＳＮを現end_of_log ＬＳＮにセットし、ヘッダ・ページに記録する必要がある。複写では、ページが短時間に逐次化されて、複写動作の実行と同様にページの修正とログが行えるようになる。

【００２５】

【実施例】

発明の方法を実行するためのＣＰＵ環境

本発明は、システム内の各ＣＰＵが、ＩＢＭのＭＶＳオペレーティング・システムを有するＩＢＭ／３６０またはＩＢＭ／３７０で構成されたＣＰＵタイプの構成中で好都合に実行できる。ＩＢＭ／３６０構成のＣＰＵは、米国特許第３４００３７１号に十分に記載されている。外部記憶装置へのＣＰＵの共用アクセスを含む構成は、米国特許第４２０７６０９号明細書に記載されている。

【００２６】また、ＭＶＳオペレーティング・システムは、ＩＢＭ刊行物ＧＣ２８－１１５０、"MVS/Extended Architecture System Programming Library:System Macrosand Facilities"、Vol.１に記載されている。ローカル・ロック管理、割込みまたはモニタによるサブシステム呼出し、タスクの記入および待機など、標準のＭＶＳまたは他のオペレーティング・システムのサービスの詳細は、省略する。これらのＯＳ処理は、当業者にはよく理解されるであろう。

【００２７】

本発明の実施態様およびトランザクション管理システム

本発明の実施例の動作を理解するには、ＴＭＳが、所与のオペレーティング・システムのもとで実行されるアプリケーションとして表されることを想起されたい。この場合、アプリケーションは、たとえば、米国特許第４４９８１４５号に記載されているようなＤＢ２型の関係データベース・システムのアプリケーションである。

【００２８】オペレーティング・システムは、その１つのタスクとして、内部記憶域と外部記憶域を資源として

含むメモリを編成する。アプリケーションから見ると、メモリと記憶域は仮想式のものであり、実要求ページングＬＲＵ階層形式の記憶域によってバックアップされる。

【００２９】ここで図１を参照すると、ＴＭＳがデータベース・システムの形で示されている。トランザクション・プロセス１２は、その実行が複数のトランザクションの並行な重なり合った実行をサポートする、アプリケーション・プログラムでもよい。プロセス１２によって実行されるトランザクションは、データベース管理システム・プログラム１４（ＤＢＭＳ）と入出力（Ｉ／Ｏ）サービスを行うオペレーティング・システム１６とを介して、データベース１３にアクセスすることができる。データ・バッファ・プール１８は、ＤＡＳＤに記憶されたデータベース１３に関するデータ用に、ＣＰＵの主メモリからＤＢＭＳ１４に割り当てられる。またＤＢＭＳ１４には、システム・ログ２１用のログ・レコードの記憶用に、主メモリからログ・バッファ２０が割り当てられる。

【００３０】ＤＢＭＳ１４は、レコード・マネージャ２２、バッファ・マネージャ２３、ログ・マネージャ２４、回復マネージャ２５、および並行性マネージャ２６を含む。

【００３１】レコード・マネージャ２２は、データベース１３のデータ構造および記憶空間を調整する。それによって、トランザクションへのレコード・レベルのアクセス、およびデータベースのロード、複写、回復などシステム・ユーティリティへのページ・レベルのアクセスが提供される。バッファ・マネージャは、データベース１３とバッファ・プール１８の間でページを移動する。ラッチ・マネージャは、バッファ・プール１８中の読み取られているまたは修正されているページに対する短期間の逐次化（共用または専用）を提供する。ログ・マネージャ２４はログ・レコードを生成し、それらをログ・バッファ２０内で番号付き順序で組み立て、ログ２１に書き込む。回復マネージャ２５は、トランザクション・レベルの回復をサポートするためにログ・レコードを利用してデータベースを戻し、同時に並行性マネージャ２６はロック・テーブル３０を介してロックを実施する。

【００３２】ここで図２を参照すると、チェックポイントおよびシステム障害に関するトランザクション・プリミティブ間での発生の関係を時間で表している。障害の発生とそれ故の再始動の際に、回復マネージャは再始動ファイルまたはその等価物から最新のチェックポイント・レコードのアドレスを獲得し、システム・ログ内でチェックポイント・レコードを探し出し、ログをその点から終りまで順方向に探索を続ける。このプロセスの結果として、資源を一貫状態に復元するために、回復マネージャは、ＵＮＤＯする必要のあるトランザクションとＲＥＤＯする必要のあるトランザクションの両方を決定することができる。

【００３３】各トランザクションは、５つのクラスのうちの１つに類別される。たとえば、タイプＴ１のトランザクションは、チェックポイント時間Ｔｃより前に完了した。タイプＴ２のトランザクションは、時間Ｔｃより前に開始し、時間Ｔｃより後でシステム障害時間Ｔｆより前に完了した。またタイプＴ３のトランザクションは、時間Ｔｃより前に開始したが、時間Ｔｆより前には完了しなかった。タイプＴ４のトランザクションは、時間Ｔｃより後で始動したが、時間Ｔｆより前に完了された。最後に、タイプＴ５のトランザクションは、時間Ｔｃより後に開始したが、時間Ｔｆまでには完了しなかった。チェックポイント時間に、修正されたページがすべてＤＡＳＤに書き込まれていたと仮定すると、トランザクションＴ２およびＴ４はＲＥＤＯを施され、トランザクションＴ３およびＴ５はＵＮＤＯを施される。

【００３４】回復マネージャは、２つのリストを始動する。ＵＮＤＯリストは最初、チェックポイント・レコードにリストされたすべてのトランザクションを含み、一方ＲＥＤＯリストは最初空である。回復マネージャは、チェックポイント・レコードから始めてログを順方向に探索する。回復マネージャが所与のトランザクションに関するＢＥＧＩＮトランザクション・レコードを見つけた場合、そのトランザクションをＵＮＤＯリストに加える。同様に、回復マネージャが所与のトランザクションに関するＣＯＭＭＩＴレコードを見つけた場合は、そのトランザクションをＵＮＤＯリストからＲＥＤＯリストに移す。

【００３５】回復マネージャは、ログ中を順方向に動作してすべてのトランザクションを再実行し、逆方向に動作してＵＮＤＯリスト内のトランザクションを取り消す。

【００３６】
好ましい実施例の方法およびプロトコル・レベルの詳細
最初に、データ・ベース中および仮想記憶域中のデータ構造と、ロギング・プロトコルおよびラッチング・プロトコルについて述べる。第２に、複写動作に関係する処理と、（必要ならば）ＩＣＢをセットするためのトランザクションの更新と、バックアップ・コピーを使用するデータ・ベースの回復について述べる。第３に、その実行が障害によって割り込まれる場合にコピーのロールバックに使用される処理について説明する。第４に、ＤＢ２のようなシステムに関連するバッファ・プールを介してではなく、ＤＡＳＤ記憶域からデータを直接複写することが可能な場合の本発明の方法の使用の仕方について述べる。最後に、共用ＤＡＳＤ記憶域を有する多重システム・トランザクション処理システムへの本発明の方法の拡張について述べる。

【００３７】データ・ベースのデータ構造
本発明の目的では、データ・ベースはデータ・ページお

よび補助構成を含む。この点に関して、ページのファイルはデータ・ベースと機能的に同等と見なす。補助構成は、割振り状況などを追跡するために、データ・ベース内にページを含む。ユーザ・データを含むページは「データ・ページ」と呼び、データ・ページの割振りおよび空間可用性状況を追跡するシステム所有のページは「スペース・マップ・ページ」（ＳＭＰ）と呼び、情報に関係するシステムを含むシステム所有のページは「ヘッダ・ページ」と呼ぶ。

【００３８】ＤＢ２の場合と同様に、大きなテーブルを、それぞれ別々のオペレーティング・システム・ファイルである多数の区画に分割することができる。各区画はヘッダ・ページと、１つまたは複数のスペース・マップ・ページと、非常に多くのデータ・ページを有することになる。アーカイブ・コピーには、データ・ページのコピーだけではなく、ＳＭＰおよびヘッダ・ページのコピーも含まれる。

【００３９】ＳＭＰはいくつかのデータ・ページの割振り状況を追跡する。またＳＭＰは、これがカバーする各データ・ページごとに１つのＩＣＢを含む。ＩＣＢの目的は、最後に複写動作が実行されて以降に、対応するデータ・ページが修正されたかどうかを追跡することである。

【００４０】ヘッダ・ページ内では、"image_copy_bit_update_ＬＳＮ"（ＩＣＢＵ＿ＬＳＮ）と呼ばれるフィールドが維持される。１つのファイルにはＩＣＢＵ＿ＬＳＮが１つだけある。ＩＣＢＵ＿ＬＳＮの目的は、それを使用することによって、データ・ページを更新するトランザクションが、そのデータ・ページに関するＳＭＰ内のＩＣＢをセットする必要があるかどうかを効率的に判定できるようにすることである。これが望ましいのは、ＩＣＢは、それが最後の複写動作以降のデータ・ページへの最初の更新である場合にだけセットする必要があるからである。変更を実行する必要がないと判定するためだけでも、ＤＡＳＤの入出力、バッファ・プール内のＳＭＰの位置決定と固定、そのラッチ、適切なＩＣＢの探索、「その固定解除」などのオーバヘッドが含まれるので、データ・ページを後で更新するためにＳＭＰにアクセスすることは避けなければならない。データ・ベースが最初にロードされた後、データ・ベースへの更新が許される前に、データ・ベースの完全コピーを取らなければならない。これはログすることなしにローディングが行われたと仮定している。これは、媒体回復が始まる前に繰り返されなければならないロードなしに、媒体回復が可能でなければならないときに必要である。完全複写動作では、ＩＣＢＵ＿ＬＳＮの初期値を、データベースを複写した後の現end_of_log ＬＳＮとして確立する。

【００４１】各アーカイブ・コピーに複写順方向ロールＬＳＮ（ＩＣＲＦ＿ＬＳＮ）が関連している。媒体回復中、これは、関連アーカイブ・コピー（最近の完全コピーおよびその後の増分式コピー）を再ロードした後でデータ・ベースを回復するためにその更新を再実行しなければならないログ・レコードを識別するためにログをそこから走査しなければならないＬＳＮである。ＩＣＲＦ＿ＬＳＮは、アーカイブ・コピーを含むファイル名や完全コピーと増分式コピーのどちらが取られたかなどの情報と共にシステム・カタログ中に記憶される。

【００４２】仮想記憶域内のデータ構造

データ・ベースが「オープン」状態にある限り、ＤＢ２のようなシステムは仮想記憶域内にそのための「データ・ベース制御ブロック」（ＤＢＣＢ）を維持する。

【００４３】更新トランザクションによって、ＩＣＢＵ＿ＬＳＮをルックアップする動作を効率的にするために、データ・ベースがオープンのときに、ＩＣＢＵ＿ＬＳＮの値が、ヘッダ・ページからＤＢＣＢ内のフィールドに複写される。その後、このフィールドが、増分式コピーの方法および手段によってのみ更新される。

【００４４】ＴＭＳは、回復のためにその更新を記述するログを使用する。ページが更新されると、システムは、更新を記述するログ・レコードに、システム規模の単調に増加するログ・シーケンス番号（ＬＳＮ）を割り当て、そのＬＳＮを"ｐａｇｅ＿ＬＳＮ"と呼ばれるフィールド内の修正されたページに記録する。

【００４５】複写動作は、単一のトランザクションとして実行され、ＩＣＢをリセットするとき、ＵＮＤＯ－ＲＥＤＯログ・レコードを書き込むものと仮定する。

【００４６】ページ・ラッチ

ＴＭＳは、単一のページ上で並行した読取り動作と更新動作または多数の並行した更新動作を逐次化するためにページ・ラッチをサポートする。共用（Ｓ）ラッチは読取り動作によって使用され、専用（Ｘ）ラッチは更新動作によって使用され、ＳモードとＸモードの間では通常の互換性規則が適用される。ページ・ラッチは、ページのファジー・コピーを行うために、複写動作によって使用される。すなわち、複写ページがコミットされていないデータを有していてもよい。

【００４７】正確判定基準

増分式コピー（ＩＣ）に関する正確さの要件は、最新の完全コピー（ＦＣ）あるいはそのＦＣの後でかつ現ＩＣの前に行われたＩＣにおいて、その効果がすでに捕捉されている更新を除き、ログ中の現ＩＣのＩＣＲＦ＿ＬＳＮ点より前に書き込まれたログ・レコードによって表されるすべてのデータ・ページ更新の効果を、現ＩＣが捕捉しなければならないことである。

【００４８】ＦＣに関する正確さ要件は、ログ中の現ＦＣのＩＣＲＦ＿ＬＳＮ点より前に書き込まれたログ・レコードによって表されるすべてのデータ・ページ更新の効果を、現ＦＣが捕捉しなければならないことである。

【００４９】バッファ・プール、システム・カタログ、および入出力の概念

バッファ・プールとは、計算または制御動作で中間結果、データ、または値を一時的に記憶するために留保されているアドレス可能な主メモリの一部分である。本発明では、他の場所で読み取り、修正し、複写する目的で、ページに関する入出力がバッファ・プール内に複写される。

【００５０】システム・カタログとは、ファイルの記憶位置に関するディレクトリである。これはまた、状況情報を含むこともできる。本発明では、これは、ファイル名、完全コピーか増分式コピーかの区別、日時、記憶される装置、および関連ＩＣＲＦ＿ＬＳＮを含む。

【００５１】入出力は、ページにアクセスするための付随的な活動を表すために使用される用語である。

【００５２】単一トランザクション管理システムにおけるアーカイブ複写

以下の諸段では、単一のＴＭＳ環境で使用する複写動作を機能的に記述する。この場合、複写動作ではデータベース・ページがＴＭＳバッファ・プールに入れられる。その後、データがＤＡＳＤから直接読み取られるとき、すなわちページがＴＭＳバッファ・プールに入れられないときに使用される複写動作について説明する。次の節では、多重システムの共用ディスク環境においてＴＭＳのために使用される複写動作について説明する。

【００５３】呼出し可能なソフトウェア機能として表される複写方法、すなわち複写動作

この節では、（ａ）複写動作、（ｂ）トランザクションによるデータ・ページの更新、および（ｃ）媒体障害後のデータ・ベースの回復を扱う。障害によって複写動作の実行が中断される場合のロールバック処理について述べる。

【００５４】アーカイブ複写は次の諸段階を含む。

【００５５】・同時に同じデータ・ベースに対して複写動作が１つだけ実行されるようにするために、システム・カタログ中に所与のデータ・ベースに関する複写動作の初期設定を登録する。

【００５６】・データ・ベースがまだオープンしていない場合に、データ・ベースをオープンし、データ・ベースのヘッダ・ページを読み取って、ＤＢＣＢ中のＩＣＢＵ＿ＬＳＮの値をセットする。これによってＤＢＣＢが生成される。データ・ベースがすでにオープンされていた場合は、ＤＢＣＢ内のＩＣＢＵ＿ＬＳＮはすでにセットされている。

【００５７】・このデータ・ベースのすべてのＳＭＰが、バッファ・プール内に確実に存在するようにする。まだバッファ・プール内にないＳＭＰについては、当然、それらのＳＭＰをバッファ・プールに送るために入出力を開始し、これらの入出力が完了するのを待つ必要がある。

【００５８】このＳＭＰの事前取り出しは、ＤＢＣＢ中のＩＣＢＵ＿ＬＳＮフィールドが可能な最大値（Ｘ'ＦＦ....ＦＦ'、次の段階を参照）を有する期間を最小にするための最適化である。

【００５９】・ＤＢＣＢ中のＩＣＢＵ＿ＬＳＮの値を原子的にＸ'ＦＦ....ＦＦ'（すなわち可能最大値）に変更する。

【００６０】・媒体回復中に使用するために、適切なシステム・カタログ中に現end_of_logのＬＳＮをＩＣＲＦ＿ＬＳＮとして記録し、その動作をログする。

【００６１】データ・ベース内の各ＳＭＰについて、以下のように初期設定を行う。

・ＳＭＰをＸ（専用）ラッチする。

・ＳＭＰ内のすべてのＩＣＢを検査して、"１"の値を有するＩＣＢを、"０"にリセットする。ＩＣでは、対応するデータ・ページは、最後の複写動作（完全または増分）以降に修正されているはずのものである。完全複写の場合は、すべてのデータ・ページが複写される。

・リセットされたすべてのＩＣＢを記述するＲＥＤＯ／ＵＮＤＯログ・レコードを、このＳＭＰ上に書き込む。ＩＣでは、どのデータ・ページを複写するかを知るために後で使用できるように、ログ・レコードのコピーを仮想記憶域に保持する。

・ＳＭＰのラッチを解除する。

【００６２】ＩＣＲＦ＿ＬＳＮの値が選択された後で、新しいＳＭＰがデータ・ベースに追加された場合、それらはこの複写動作の一部分として検査する必要はないことに留意されたい。

【００６３】そのような各ＩＣＢｉについて、以下のことを行う。

【００６４】ｉ番目のデータ・ページをＳ（共用）ラッチする。

【００６５】ｉ番目のデータ・ページをアーカイブ・コピー・ファイルに複写する。

【００６６】ｉ番目のデータ・ページをラッチ解除する。

【００６７】ＦＣについては、このＳＭＰでカバーされるすべてのデータ・ページが複写される。

【００６８】これらの動作を極めて効率的にし、かつ入出力のコストおよび遅延を減少させるために、バッチ式入出力（すなわち、１回の入出力呼出しで１ページよりも多く読む）および所望のデータ・ページの事前取出しを使用することができる。また並列処理を用いて、また複写動作の経過時間をさらに減少させることもできる。

【００６９】関係する動作には以下のものが含まれる。

・現end_of_log ＬＳＮを得る。

・ＤＢＣＢおよびヘッダ・ページ内のＩＣＢＵ＿ＬＳＮフィールドを、現end_of_log ＬＳＮの値に更新する。

・ＲＥＤＯ＿ｏｎｌｙログ・レコードを使用して、その更新をヘッダ・ページにログする。

【００７０】アーカイブ複写トランザクションをロールバックすべき場合も、古いＩＣＢＵ＿ＬＳＮに戻るので

はなく、新しいＩＣＢＵ＿ＬＳＮを保持するのが正しいはずである。これが、ＲＥＤＯ／ＵＮＤＯログ・レコードではなく、ＲＥＤＯ＿ｏｎｌｙログ・レコードを使用する理由である。これに関連して、この方法は、共用ＤＡＳＤ環境でヘッダ・ページ更新のＵＮＤＯの何らかの特別な扱いを回避する。

【００７１】データ・ベース内のそれぞれのＳＭＰについて、以下のことを行う。

【００７２】ＩＣについて、そのＳＭＰに関して前に書き込まれたログ・レコードの、キャッシュされたバージョンを検査することによって、その動作によって（"０"に）リセットされたＩＣＢを決定する（ＩＣＢｉはｉ番目のデータ・ページと関係する）。

【００７３】複写動作の終了時に（すなわち、すべての関連データ・ページが複写された後で）、次のことを行う。

【００７４】この複写動作が完了しており、したがってこのイメージ複写トランザクションがコミットされた後は、このデータ・ベースに関するイメージ複写動作のその後の呼出しが遅延なしで実行を開始できることに留意されたい。

【００７５】複写トランザクションをコミットする。

【００７６】複写動作の実行中に並行更新が許されない場合は、その方法は上記の方法の一部分である。ＩＣＲＦ＿ＬＳＮの値は、必要なページをすべて複写した後で選択される。この複写動作には、ページのラッチは必要でない。

【００７７】ページを更新するトランザクションのための方法または論理

更新トランザクションは、データ・ページを更新する間に、そのデータ・ページに関するＩＣＢをＳＭＰ中で（"１"に）セットする必要があるかどうかを判定しなければならない。これは、（この更新を行う前に）更新されるページの現ｐａｇｅ＿ＬＳＮをＩＣＢＵ＿ＬＳＮと比較することによって行われる。この検査により、データ・ページが複写動作の最後の実行以降にすでに修正されているか、それともその時以降で最初の更新であるかを判定する。

【００７８】更新トランザクションが従う方法または論理は、次のとおりである。

【００７９】データ・ページをＸラッチする。

【００８０】現ｐａｇｅ＿ＬＳＮがＤＢＣＢ内のＩＣＢＵ＿ＬＳＮ（更新前）よりも小さいときは、次のことを行う。

【００８１】このデータ・ページをカバーしＳラッチするＳＭＰにアクセスする。

【００８２】複写動作が実行を開始したすぐ後にＩＣＢＵ＿ＬＳＮがX'ＦＦ....ＦＦ'にセットされるので、このデータ・ベースに対するすべての更新がこの動作の実行中に関連ＳＭＰを訪問することになる。

【００８３】ＩＣＢがまだセットされていない場合は、それをセットし、ＲＥＤＯ＿ｏｎｌｙログ・レコードを使用する動作をログする。ＳＭＰ上ではＳラッチしか獲得されていないので、複数のトランザクションがＳＭＰを並行して更新中であることがあり得る。Ｓラッチ処理は、そのようなホット・スポット・ページ上での並行性を改善するために行われる。その結果、ｐａｇｅ＿ＬＳＮフィールドのみならずＩＣＢに対する更新も慎重に行わなければならない（たとえば、ｐａｇｅ＿ＬＳＮフィールドが別のトランザクションによって高い値に更新されている場合はそれを更新してはならず、比較交換の方法または論理を使用すべきである）。

【００８４】次の場合には、ＩＣＢはすでにセットされていることがあり得る。

【００８５】ＤＢＣＢ内のＩＣＢＵ＿ＬＳＮがX'ＦＦ....ＦＦ'にセットされた状態で複写動作が実行中に、データ・ページが１回または複数回更新された。その間にＳＭＰを何度も訪問することを回避する１つの方法は、ＩＣＢＵ＿ＬＳＮがX'ＦＦ....ＦＦ'の間に、データ・ページのＩＣＢがすでにセットされたことを示すことのできるフラグを、そのページに関するバッファ・マネージャのページ制御ブロック（ＰＣＢ）内に立てることである。複写動作は、データ・ページ上にラッチを保持している間、そのフラグをリセットする。そのフラグは、そのデータ・ページがある更新から別の更新にキャッシュされたままである場合、すなわちそのページに関するバッファ・スロットがスチールされていない場合だけ役立つ。ページが置き換えられ再びキャッシュされた場合、ＰＣＢフラグがＩＣＢがセットされていないことを意味するように初期設定される。

【００８６】ＳＭＰ更新のためにＲＥＤＯ専用レコードとしてログ・レコードを書く理由は、そのデータ・ページを更新したトランザクションがロールバックすべき場合でも、ＳＭＰ内でＩＣＢの値が（"１"に）セットされたままでなければならないからである。これは、別のトランザクションによる同じデータ・ページへのその後の更新では、ｐａｇｅ＿ＬＳＮが最初の更新によってＩＣＢＵ＿ＬＳＮよりも大きい値にセットされているはずだから、ＩＣＢはセットされているはずだと仮定するからである（以下参照）。

【００８７】ＳＭＰのラッチを解除する。

【００８８】ＤＰを更新し、更新をログし、ＤＰのｐａｇｅ＿ＬＳＮを、書き込まれたばかりのログ・レコードのＬＳＮにセットする。

【００８９】ＤＰをラッチ解除する。

【００９０】ＩＣＢＵ＿ＬＳＮが最大値にセットされたままである時間を減少させるための最適化

バッファ・プール内にその時キャッシュされているページのリストを維持するＤＢ２などのＴＭＳ内では、ＩＣＢＵ＿ＬＳＮが最大値になっている時間を減少させるこ

とができる。これは以下のようにして実施される。

【００９１】ＩＣＢＵ＿ＬＳＮを最大値にセットした後、変更されたページのリスト上の各ページをＳラッチする。

【００９２】ＩＣＢＵ＿ＬＳＮを現ログ終端（非最大値）にセットする

【００９３】ページを複写する。

【００９４】この最適化では、ＩＣＢＵ＿ＬＳＮを非最大値にセットした後でページがアクセスされ複写されるので、ＩＣＢＵ＿ＬＳＮが最大値にセットされている時間が減少する。

【００９５】ＩＣＢをセットする必要がある場合は、ＩＣＢがセットされＩＣＢセット動作がログされた後で、対応するＤＰの更新のロギングを行うことが重要であることに注意されたい。順序を逆にすると媒体回復が正しくなくなることがある。

【００９６】新しく割り振られたページについては、ＩＣＢは常にセット（"１"）される。

【００９７】複写動作実行中の障害

複写動作の実行が、何らかの障害によって中断される場合は、たとえば複写動作によって書き込まれたログ・レコードを使用する通常のロールバックの方法または論理により、その設定が複写動作によって修正されたＩＣＢの古い設定が復元される。他のＩＣＢの値が、ロールバック法によって、それらの元の値を含むように修正されない（すなわち、どのＩＣＢにもロールバック中に"０"が割り当てられない）ことが重要である。また、現複写実行に関するシステム・カタログ内の情報（たとえば、ＩＣＲＦ＿ＬＳＮの値）は削除されることになる。

【００９８】媒体障害後の回復

媒体回復のための方法または論理では、最後に完了した複写動作のＩＣＲＦ＿ＬＳＮの値からログ走査を始める必要がある。この走査は、最新のＦＣおよび後続のすべてのＩＩＣから時間順にデータが複写された後で開始される。

【００９９】実例

ここで図３を参照すると、本発明で使用される論理ファイルおよびページ編成が示されている。さらに詳細には、ファイルまたはデータ・ベースは、ヘッダ・ページ、スペース・マップ・ページ（ＳＭＰ）、および複数のデータ・ページ（Ｄページ１～４）を含む。任意に「ファイルＡ」と名付けたデータが、ＣＰＵ制御のロード動作の制御下で、ＤＡＳＤ外部記憶装置内のデータベース１３にロードされる。ログ・スペースとログ・オーバヘッドを節約するために、ロードはログを中断した状態で実行される。また、ロード動作の実行中、並行更新は許されない。

【０１００】この動作の一部分として、各ページにページ番号が連続して割り当てられる。ヘッダ・ページの番号はページ０、ＳＭＰはページ１、Ｄページ１～４はページ２～５と番号が付けられる。ここでは、ページ番号は単にローカル・ページ・シーケンスを定義するものにすぎず、ｐａｇｅ＿ＬＳＮと混同してはならない。後者は各ページに埋め込まれたポインタであり、ログされた後でＤＡＳＤ記憶装置に書き込まれたとき、最新のページ修正を含むログ中のＲＥＤＯ／ＵＮＤＯレコードの位置を示す。最初に、ｐａｇｅ＿ＬＳＮが０にセットされる。同様に、すべてのＤページに関するＳＭＰ内のＩＣＢが０にセットされ、ヘッダ・ページ内のＩＣＢＵ＿ＬＳＮも同様である。最後に、ファイルＡに対するバックアップは存在しないので、システム・カタログ内にエントリはない。

【０１０１】図３から理解されるように、システム・カタログは、ファイルＡおよび他のファイルの完全コピーまたは増分式コピーに関する位置ポインタと他のアーカイブ・コピーまたはバックアップ・コピー情報を幅広く含む。一般に、そのようなアーカイブ・コピーは、自動テープ・ライブラリなどの補助記憶装置に記憶される。バックアップ・コピーの回復使用には、テープ・ライブラリへのアクセスとそのテープ・ライブラリからＤＡＳＤ記憶装置へのステージングが必要である。

【０１０２】ここで図４を参照すると、初期ロードの後データ・ページに許される修正の前に、ファイルＡの完全コピー（ＦＣ）を取った後のファイルＡおよびシステム・カタログの状態が示されている。図４で、それがＦＣ完了後の現end_of_logであったので、ＩＣＢＵ＿ＬＳＮおよびＩＣＲＦ＿ＬＳＮが１００にセットされていることに留意されたい。ヘッダ・ページは、その後１１０のＬＳＮを有するログ・レコードを書き込むことにより、ＩＣＢＵ＿ＬＳＮのこの値を記録する方法で更新される。すなわち、ヘッダ・ページが、現ＩＣＢＵ＿ＬＳＮ値１００を記憶しているログの位置がログ位置１１０である。

【０１０３】ここで、図５を参照すると、ファイルＡのＦＣが１９９２年５月１日２３時５０分に取られ、テープ・ライブラリ・アドレスＴ００１に記憶された、システム・カタログ・エントリが示されている。ＦＣが取られた後でＤＡＳＤ　Ｄ００１（図示せず）が失われる場合、ファイルＡに関するＤＢＭＳ回復動作は次の諸段階を含むことになる。

【０１０４】（ａ）システム・カタログからのファイルＡの最近の完全コピーに関する現ログ終端（任意にＥＮＤ＿ＬＳＮと名付ける）とテープ位置を確認する。

【０１０５】（ｂ）ファイルＡのＦＣバックアップを、テープ位置Ｔ００１からＤＡＳＤ　Ｄ００２（図示せず）にロードする。

【０１０６】（ｃ）システム・カタログからのＦＣの後でファイルＡの増分式コピー（ＩＩＣ）を確認し、それに応じて更新する。（段階（ｃ）のこの場合には、ＦＣの後でファイルＡのＩＩＣがまだ取られていない）

【0107】（d）それが最新のバックアップなので、システム・ログをICRF_LSN＝100に位置決めし、page_LSN<ログ・レコードLSNの場合は、DASD D002にあるファイルAのページの更新に関するログ・レコードを適用する（この例では、ICRF_LSN＝100からEND_LSNである）。

【0108】段階（d）中のログ走査の回数を減らすために、DBMSは修正されたシステム・カタログ内に配置された別のディレクトリを利用することもできる。この別のディレクトリは、更新のためにファイルAが開かれている間のログ・レンジからなることが好ましい。この例で「SYSLGRNG」と記されたこのディレクトリ（図示せず）は、ヘッダ・ページの更新に関する100から110のLSNを表す、ファイルAのエントリを有する。その後、回復動作が完了する。

【0109】ここで、図6ないし図10を参照をすると、選択されたデータ・ページ（ページ1とページ3）のトランザクションによる修正が示されている。ヘッダ・ページがICBU_LSN＝100とDページ1およびDページ3のpage_LSN＝0を持つことを思い起こされたい。

【0110】Dページ1の更新では、図6に示す通り、ファイルAに関するICBU_LSN＝100がまたデータ・ベース制御ブロック（DBCB）内に記録される。Dページ1のpage_LSN0はICBU_LSN100よりも小さいので、そのとき更新方法で、Dページ1を更新する前に、SMP内のICB1をセットしなければならない。ICB1の変更と、変更がログ（LSN212）内に記録されるLSNは両方とも、図7に示したようにSMP上に示される。次に、Dページ1の値がA1からB1に変えられ、LSN215に記録されたログになる。これらの変更は、図8の更新されたDページ1に示される。

【0111】Dページ3の更新では、Dページ3のpage_LSN0はICBU_LSN100よりも小さいので、図9からわかるように、SMP内のICB3はDページ3の更新前に1にセットされる。これに続いて、図10に示すように、Dページ3は値A3からC3に更新され、これがLSN326にログされたLSNもDページ3でマークされる。

【0112】ここで図11を参照すると、ファイルAが、更新されて、SMPおよびページ更新の変更を含むDASDに書き込まれたものとして示されている。

【0113】ここで図12ないし図23を参照すると、図11に描かれたファイルAのICが示されている。図12および図13は、DBCBとSMPによるIC動作に関するファイルAの初期状態を示す。最初の段階は、バッファ・プール内にSMPを読み込むことである。次に、図14に示すように、ICBU_LSNが16進表記で最大値"FFFF....FF"に変更される。これに続いて、ICRF_LSNが、現end_of_log＝463にセットされる。ここで、図15に示すように、システム・カタログが、ICが取られていること示す新しいエントリ、日時、テープ内のICの位置、およびICRF_LSNで更新される。

【0114】次に仮想記憶域内のSMPを更新することが望まれる。これは、SMPをラッチして、それを図16に示したように変更する必要があることを意味する。すなわち、SMPをラッチした後、"1"であるICBが"0"にリセットされ、図17に示したようなUNDOレコードがログに書き込まれ、その後にSMPがラッチ解除される。

【0115】IC複写動作は、そのICBが"1"から"0"にリセットされた、ヘッダ・ページ、SMPページおよびDページを複製する段階を含む。Dページを含む各ページを、ラッチし、テープT002に複写し、次いでラッチ解除する必要がある。図18に示すように、テープ上の各ページは、Dページ1のそれと同様に新しい。完璧を期して、図21に示すように、ICはDページ3のテープT002へのコピーを作成させる。最後の段階では、ICBU_LSNに関する現end_of_log値を獲得し、それをDBCBと更新されたヘッダ・ページの両方に複写し、それをヘッダ・ページに関してログし、図22および図23に示すヘッダ・ページ内にpage_LSNを置く。

【0116】さらに、再び図12ないし図23を参照して、図18に示すように、テープT002に複写された直後に、別のトランザクションがDページ1を更新しようとするものとする。図16に示すように、増分式に複写されるページのセットの決定の一環として、すべてのICBが"0"にセットされていることを想起されたい。したがって、Dページ1への現更新にはICB1をセットすることが必要である。ここで、Dページ1の場合、page_LSN＝215である。これは、最大値にセットされたICBU_LSNよりも小さい。したがって、図19および図20に示すように、ICB1はSMP中で1にセットされ、Dページ1が更新され、両方のページがログされ、LSNが当該のページに記録される。

【0117】重要なことであるが、ICBU_LSNの値が100のままであった場合、Dページ1の更新によって、ICB1が"1"にセットされることにはならないはずである。したがって、後続のIICは、値"C1"を有するDページ1を複写しないことになる。後続のICが完了すると、ICRF_LSNは504よりも大きくなる。次のIC後にファイルAが失われる場合、504よりも大きいICRF_LSNを使用したバックアップ・コピーの復元により、LSN504の更新が失われることになる。これにより、ICの間に新しい非最大値が確立されつつある間ICBU_LSNを最大値にセット

することの必要性が説明できる。コピー・セット中のすべてのページをラッチし、複写し、ラッチ解除することを含めてＩＣを完了した後で、ＩＣＢＵ＿ＬＳＮを非最大値にセットすることが必要であり、そうしないと、対応するＩＣＢをセットするために、あるいはＩＣＢがすでにセットされているかどうかを調べるために、ファイル内のページへのあらゆる更新でＳＭＰへのアクセスが起こることになる。これは無駄だと考えられる。

【０１１８】この例では、Ｄページ３はまだ複写されておらず、したがって、ＩＣＢＵ＿ＬＳＮの新しい非最大値は確立されていない。Ｄページ１を更新したトランザクションは、そのｐａｇｅ＿ＬＳＮを比較するためにＩＣＢＵ＿ＬＳＮの最大値を使用した。

【０１１９】ＩＣＢＵ＿ＬＳＮの新しい非最大値をセットする際の他の要因

ＩＣＢＵ＿ＬＳＮの新しい非最大値をセットする際に考慮すべき要因が他にもある。ここでは、比較のために古いＩＣＢＵ＿ＬＳＮを使用し、したがってＩＣＢビットの設定を迂回した可能性のある更新トランザクションがなく、その新しいｐａｇｅ＿ＬＳＮがＩＣＢＵ＿ＬＳＮよりも大きいことが必要である。そのようなことは、更新トランザクションがそのｐａｇｅ＿ＬＳＮをＩＣＢＵ＿ＬＳＮの古い値と比較した後で、そのトランザクションがオペレーティング・システムによって待機中の活動または予定の活動から取り除かれる場合に起こり得る。これにより、ＩＣの間にＩＣＢＵを最大値にセットすることによって避けようとした状況が生じることになる。すなわち、値が最大値ではなかった場合、ページへのその後の更新が、そのＩＣＢがセットされなかったため、次の複写動作で捕捉されないことになる。

【０１２０】この解決法では、ページをＳ（共用）ラッチする必要がある。これにより、更新トランザクションがページへの更新を完了したことが保証される。すなわち、ラッチによって原子的動作がもたらされる。ページの一貫性のあるコピーをとるために、Ｓラッチが必要とされることに留意されたい。

【０１２１】ただ１つのＩＣＢＵ＿ＬＳＮの値が選択されるので、そのような選択は、コピー・セット中のすべてのページをＳラッチした後でだけ行われる。

【０１２２】ここで図２４および図２５を参照すると、テープ記憶装置にアーカイブ保存された、完全複写および増分式複写動作後のファイルＡが示されている。現コピーを記憶するＤＡＳＤが障害を起こしたために、ファイルＡの現状態を回復する必要があるものと仮定する。この場合、図１５に示したシステム・カタログに従って、最新の完全コピーとその後の増分式コピーの位置が確立される。それはＩＣＲＦ＿ＬＳＮ＝４６３である。これは、そのときそれぞれテープ・ライブラリ・アドレスＴ００１とＴ００２にアーカイブ保存されたページ番号によって完全コピーおよび増分式コピーをマージすることに基づいて、ファイルＡを回復することを意味する。

【０１２３】ここで、図２６を参照すると、記憶テープからのファイルＡのページ・マージされたコピーが示されている。ページの状態を順方向に変化させるために、ログがＬＳＮ４６３から走査され、ＬＳＮ４９０でＳＭＰに、ＬＳＮ５０４でＤページ１に、５３０でヘッダ・ページに更新を適用する。これにより、図２７に示したような最終ファイル状態になる。

【０１２４】ここで、図２８および図２９を参照すると、少なくとも１ページへの複数の更新の後のファイルおよび構成の状態が示されている。この場合、トランザクションはＤページ４と次いでＤページ１を更新する。

【０１２５】Ｄページ４については、ｐａｇｅ＿ＬＳＮ（０）がＩＣＢＵ＿ＬＳＮ（５１９）よりも小さいので、ＩＣＢ４はＳＭＰ中で１にセットされ、ＳＭＰはｐａｇｅ＿ＬＳＮ＝５７０にログされ、Ｄページ４はＢ４に更新されｐａｇｅ＿ＬＳＮ＝６２０にログされることになる。

【０１２６】Ｄページ１については、ｐａｇｅ＿ＬＳＮ（５０４）がＩＣＢＵ＿ＬＳＮ（５１９）よりも小さいので、ＳＭＰはアクセスされるが、ＩＣＢ１がすでに１にセットされているので、ＳＭＰへの更新は必要でない。次に、Ｄページ１が値Ｄ１に更新され、ｐａｇｅ＿ＬＳＮ＝６３０にログされる。

【０１２７】ここで、ＩＣを取るべき場合、ＳＭＰに従ったコピー・セットはＤページ１およびＤページ４となる。

【０１２８】本発明の上記その他の拡張は、頭記特許請求の範囲に記載されるその趣旨および範囲から逸脱することなしに行うことができる。

【０１２９】

【発明の効果】本発明の実施により、ログ・ベースの段階式コミットＴＭＳのデータベース中の修正されたページの増分式複写に伴うオーバヘッドを減少させるための方法および手段を提供することができ、アプリケーションの実行を妨げないログ・ベースの段階式コミットＴＭＳ内で修正可能なページのデータベースをアーカイブ保存するための方法および手段を提供することができ、修正可能なページのデータベースへのアクセスを共有する複数ログ・ベース段階式コミット・トランザクション管理システムで動作可能な方法および手段を提供することができる。

【図面の簡単な説明】

【図１】従来技術によるログ・ベースＴＭＳの論理編成を示す図である。

【図２】従来技術による、ＲＥＤＯおよびＵＮＤＯに関する並列トランザクションと、ログ・ベースの段階式コミット・トランザクション管理システム内での障害の発生との関係を示す図である。

【図３】データベースを表す小さなファイルと、分類されたシステム構造およびデータ・ページを格納するシステム・カタログの論理編成を示す図である。

【図４】データベースの初期ロード中にログが中断されたものと仮定して、始めにデータをロードし必要なその完全コピーを取った後のファイルを示す図である。

【図５】完全バックアップのためのシステム・カタログのエントリを示す図である。

【図６】様々なデータ・ページへのトランザクションの修正と様々なＳＭＰおよびＬＳＮの変更を示す図である。

【図７】様々なデータ・ページへのトランザクションの修正と様々なＳＭＰおよびＬＳＮの変更を示す図である。

【図８】様々なデータ・ページへのトランザクションの修正と様々なＳＭＰおよびＬＳＮの変更を示す図である。

【図９】様々なデータ・ページへのトランザクションの修正と様々なＳＭＰおよびＬＳＮの変更を示す図である。

【図１０】様々なデータ・ページへのトランザクションの修正と様々なＳＭＰおよびＬＳＮの変更を示す図である。

【図１１】図６ないし図１０に示した変更を統合したバージョンを示す図である。

【図１２】本発明による増分式複写動作に関連する状態およびページの変更を示す図である。

【図１３】本発明による増分式複写動作に関連する状態およびページの変更を示す図である。

【図１４】本発明による増分式複写動作に関連する状態およびページの変更を示す図である。

【図１５】本発明による増分式複写動作に関連する状態およびページの変更を示す図である。

【図１６】本発明による増分式複写動作に関連する状態およびページの変更を示す図である。

【図１７】本発明による増分式複写動作に関連する状態およびページの変更を示す図である。

【図１８】本発明による増分式複写動作に関連する状態およびページの変更を示す図である。

【図１９】本発明による増分式複写動作に関連する状態およびページの変更を示す図である。

【図２０】本発明による増分式複写動作に関連する状態およびページの変更を示す図である。

【図２１】本発明による増分式複写動作に関連する状態およびページの変更を示す図である。

【図２２】本発明による増分式複写動作に関連する状態およびページの変更を示す図である。

【図２３】本発明による増分式複写動作に関連する状態およびページの変更を示す図である。

【図２４】完全複写動作および増分式複写動作がアーカイブ保存された後のテープ記憶装置上のファイルＡのコピーを示す図である。

【図２５】完全複写動作および増分式複写動作がアーカイブ保存された後のテープ記憶装置上のファイルＡのコピーを示す図である。

【図２６】図２４および図２５のＦＣおよびＩＣとログ・レコードとを用いた回復後のファイルＡを表す図である。

【図２７】図２４および図２５のＦＣおよびＩＣとログ・レコードとを用いた回復後のファイルＡを表す図である。

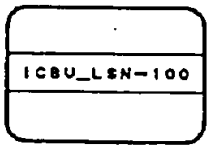
【図２８】少なくとも１つのページへの多重更新の後のファイルおよび構成の状態を示す図である。

【図２９】少なくとも１つのページへの多重更新の後のファイルおよび構成の状態を示す図である。

【符号の説明】

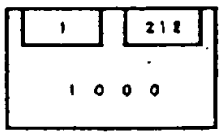
１２　トランザクション・プロセス
１３　データベース
１４　データベース管理システム・プログラム
１６　オペレーティング・システム
１８　バッファ・プール
２０　ログ・バッファ
２１　システム・ログ
２２　レコード・マネージャ
２３　バッファ・マネージャ
２４　ログ・マネージャ
２５　回復マネージャ
２６　並列マネージャ
３０　ロック・テーブル

【図６】



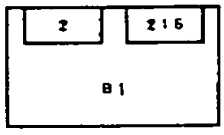
ファイル "A" 用のDBCB

【図７】



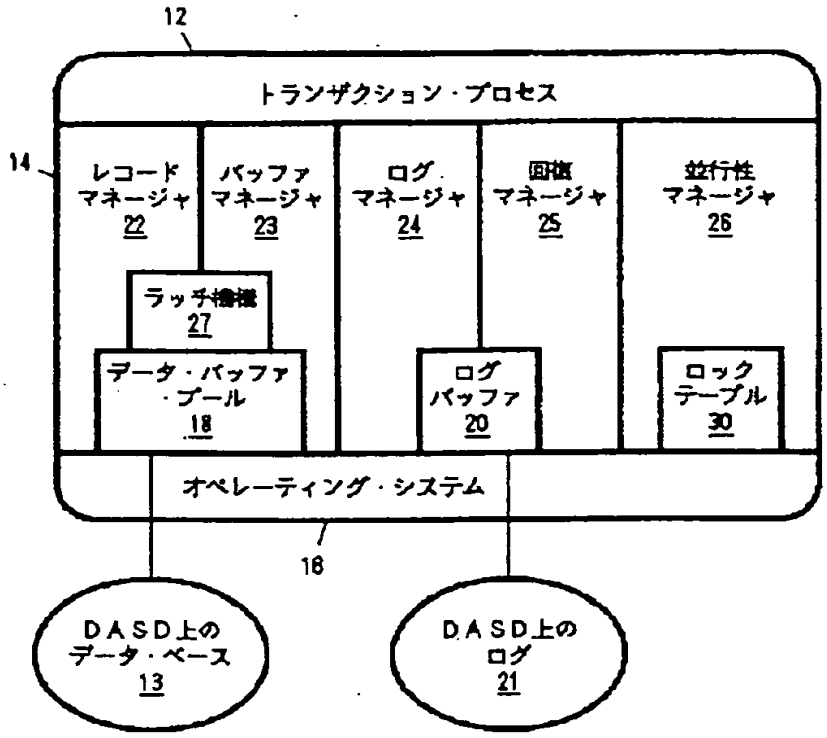
Dページ1に関してセットされたICB用のSMPの更新

【図８】



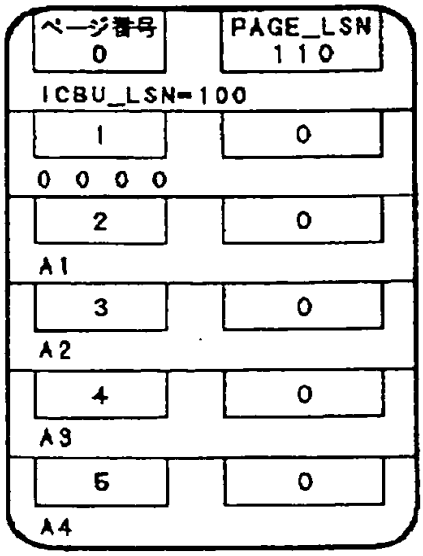
Dページ1を値A1からB1に更新する

【図１】



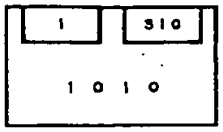
ログ・ベースの段階式コミット・トランザクション管理システム（従来技術）

【図４】



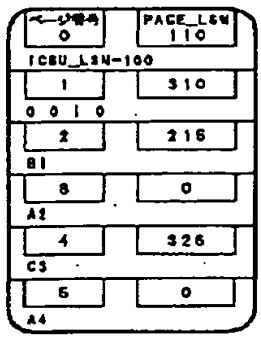
初期のロードを行い完全イメージコピーされた後のファイル内のページ

【図９】



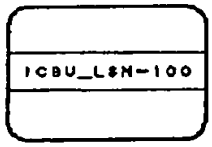
Dページ3に関してセットされたICB用のSMPの更新

【図１１】



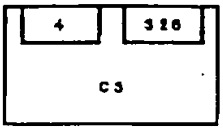
Dページ1とDページ3が更新された後のファイルA

【図１２】



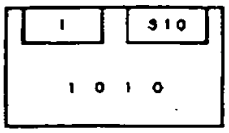
ＩＩＣ開始時のDBCB

【図１０】



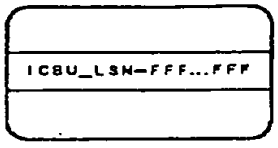
Dページ3を領域A3からC3に更新する

【図１３】



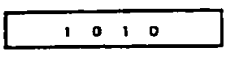
ＩＩＣ開始時のSMP

【図１４】



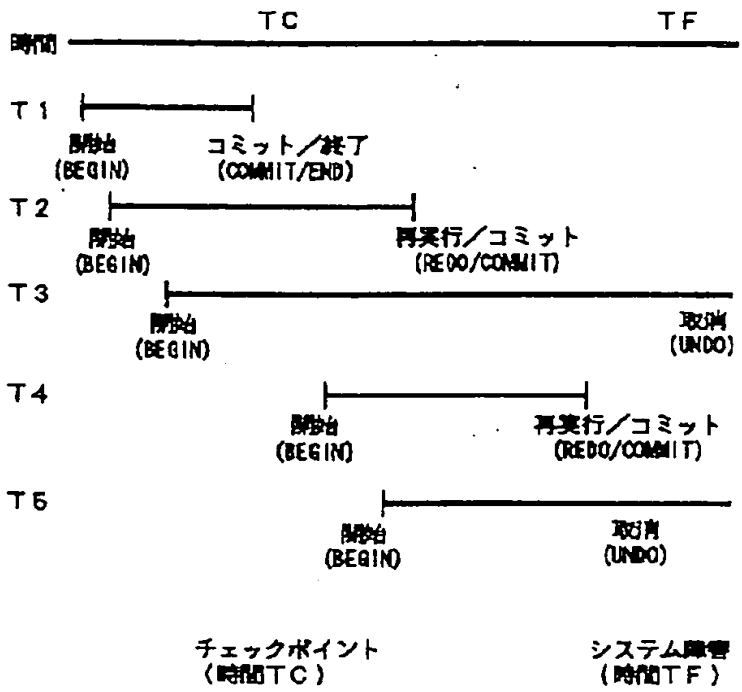
最大値にセットされたICBU_LSNを有するDBCB

【図１７】



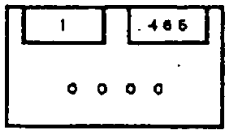
ログ取得レコード

【図２】



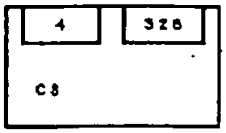
ログ・ベースの段階式コミット・トランザクション管理システムにおけるトランザクション，進行，チェックポイント，障害，再実行および取消の関係（従来技術）

【図１６】



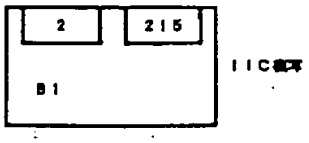
仮想記憶域内のSMP

【図２１】



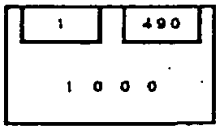
テーブルT002に属するDページ3

【図１８】



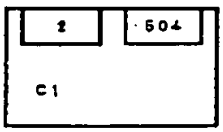
Dページ1

【図１９】



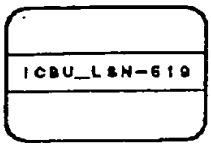
IC後のDページ1更新に応答するSMP

【図２０】



IC領域に修正されたDページ1

【図２２】



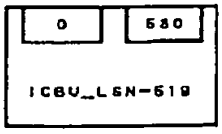
ログ領域にセットされたICBU_LSNを有するDBCB

【図５】

ファイル名	日付	時間	完全または増分式	ファイル・コピーのテープ・アドレス	ICRF_LSN
A	01592	2350	完全	T001	100

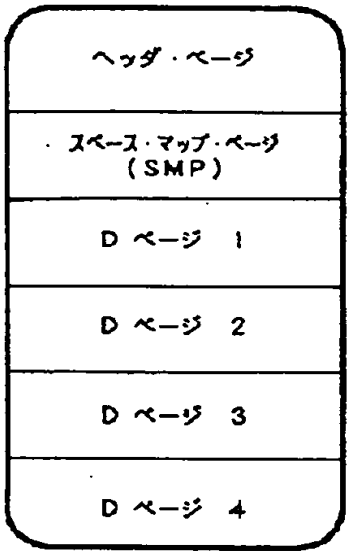
完全イメージ・コピー後のシステム・カタログのエントリ

【図２３】

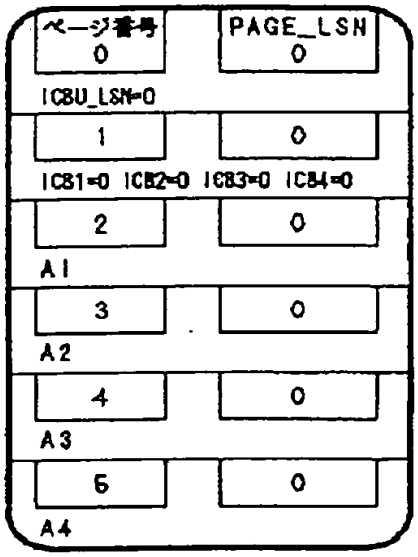


新しいICBU_LSNの許容最大値を有する修正されたヘッダ・ページ

【図３】

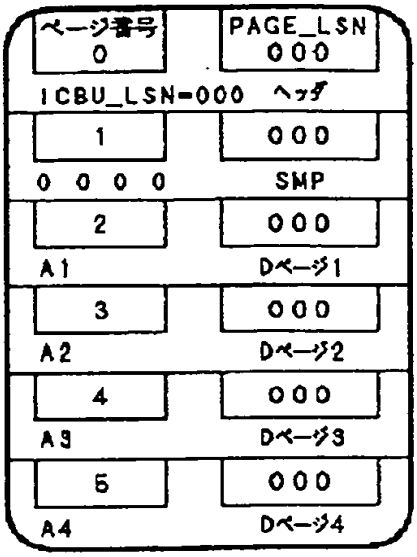


ファイル "A"



ファイル "A" 内のページ

【図２４】



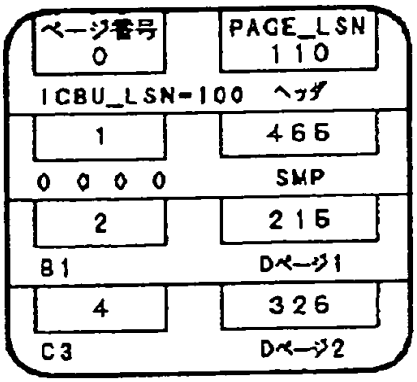
テープＴ００１上のＦＩＣ

ファイル名	日 付	時 間	完全または増分式	ファイル・コピーのテープ・アドレス	ICRF_LSN

システム・カタログ

論理ファイル，ページおよびシステム・カタログの構成

【図２５】



テープＴ００２上のＩＩＣ

【図１５】

ファイル名	日付	時間	完全または増分式	ファイル・コピーのテープ・アドレス	ICRF_LSN
A	010592	2350	FIC	T001	100
A	010682	2200	IIC	T002	483

システム・カタログ

【図２６】

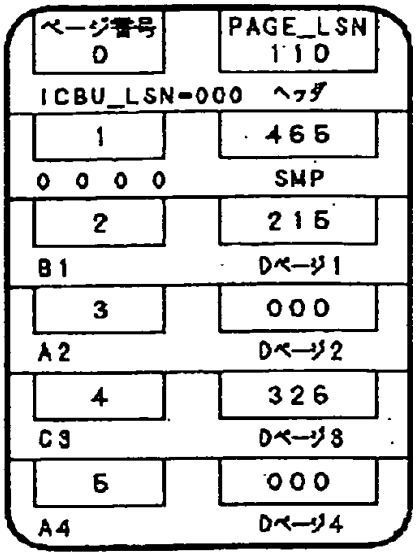
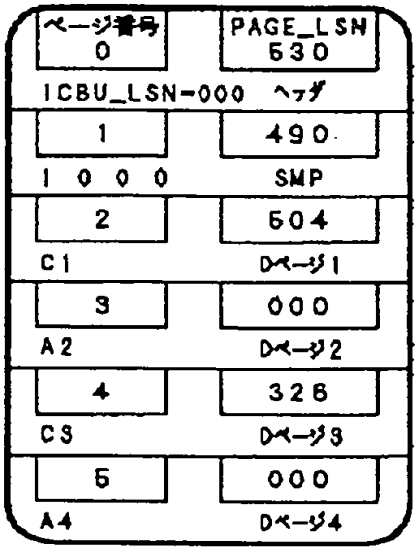


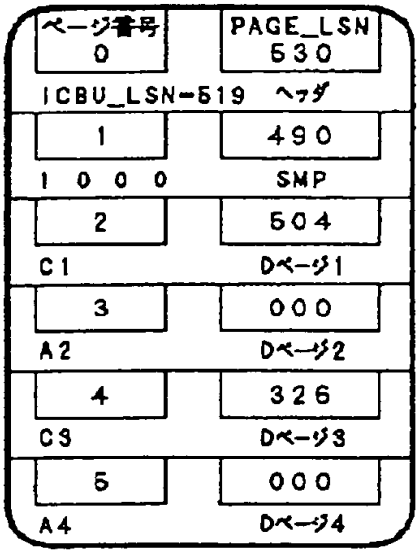
図２４と図２５のＦＣとＩＣに基づく回復

【図２７】

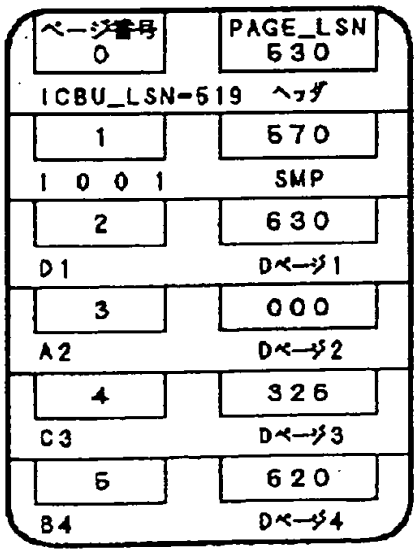


ICRF_LSN=483 から順方向にロールした後の図２６のファイル

【図２８】



【図２９】



フロントページの続き

(72)発明者　インデルパル・エス・ナラング
アメリカ合衆国95070　カリフォルニア州
サラトガ　セラ・オークス・コート13778